

CUADERNO DE TRABAJO N°3-2026

DEL MARTILLO AL AGENTE, DE ESCUDO A AMENAZA: LA INDUSTRIALIZACIÓN DEL CIBERATAQUE Y LA RECONFIGURACIÓN DEL PODER EN LA ERA DE LA INTELIGENCIA ARTIFICIAL



Academia Nacional
de Estudios Políticos
y Estratégicos

www.anepe.cl



CUADERNOS DE TRABAJO es una publicación orientada a abordar temas vinculados a la Seguridad y Defensa a fin de contribuir a la formación de opinión en estas materias.

Los cuadernos están principalmente dirigidos a tomadores de decisiones y asesores del ámbito de la Defensa, altos oficiales de las Fuerzas Armadas, académicos y personas relacionadas con la comunidad de defensa en general.

Estos cuadernos son elaborados por investigadores, académicos y colaboradores del CIEE de la ANEPE, pero sus páginas se encuentran abiertas a todos quienes quieran contribuir al pensamiento y debate de estos temas.

Recordamos a los autores que el Cuaderno de Trabajo está comprometido con la publicación de artículos originales e inéditos que difundan conocimiento actualizado en materias de seguridad, defensa y ciencias sociales afines, con el fin de aportar y transferir, con el propósito fundamental de aportar al debate académico múltiples enfoques que enriquezcan el análisis, la reflexión y la interpretación en torno a los temas disciplinares propios de la seguridad, la defensa y las ciencias sociales.



Antes de imprimir este Cuaderno, piense en el medio ambiente.

CUADERNO DE TRABAJO DEL CENTRO DE INVESTIGACIONES Y ESTUDIOS ESTRATÉGICOS es una publicación electrónica del Centro de Investigaciones y Estudios Estratégicos de la Academia Nacional de Estudios Políticos y Estratégicos y está registrada bajo el **ISSN 0719-4110 Cuad. Trab., - Cent. Estud. Estratég.**

Dirección postal: Avda. Eliodoro Yáñez 2760, Providencia, Santiago, Chile.

Sitio Web www.anepe.cl. Teléfonos (+56 2) 2598 1000, correo electrónico ciee@anepe.cl

Todos los artículos son de responsabilidad de sus autores y no reflejan necesariamente la opinión de la Academia.

Autorizada su reproducción mencionando el Cuaderno de Trabajo y el autor.

CONSEJO EDITORIAL

Director

Cristian Château Magalhães

Magíster en Ciencias Militares y Profesor de Academia en Historia Militar, Estrategia y Táctica en Operaciones (ACAGUE).

ORCID: <https://orcid.org/0000-0001-7144-8133>

Editor Ejecutivo

Fulvio Queirolo Pellerano

Magíster en Ciencia Política, Seguridad y Defensa (ANEPE). Magíster en Ciencias Militares y Profesor de Academia en Historia Militar y Estrategia (ACAGUE). Doctor (c) en Seguridad Internacional (UNED, España).

ORCID: <https://orcid.org/0000-0001-6837-0962>

Consejeros de Editorial

Carlos Ojeda Bennett

Doctor en Ciencia Política, Universidad Paris V. Magíster en Ciencias Militares (ACAGUE). Profesor en Estudios de Seguridad y Manejo de Crisis (ANEPE).

ORCID: <https://orcid.org/0000-0001-8260-5510>

Jorge Sanz Jofré

Doctor en Desarrollo Local y Territorio por la Universitat Jaume I. y Universitat de Valencia, España. Magíster en Ciencias Militares y Profesor de Academia en Geopolítica y Estrategia (ACAGUE).

ORCID: <https://orcid.org/0000-0003-2767-0701>

Consejero Externo

Jorge Gatica Bórquez

Doctor en Estudios Americanos (USACH). Magíster en Ciencia Política (PUC). Magíster en Ciencias Militares (ACAGUE).

ORCID: <https://orcid.org/0000-0003-1596-5588>

DEL MARTILLO AL AGENTE, DE ESCUDO A AMENAZA: LA INDUSTRIALIZACIÓN DEL CIBERATAQUE Y LA RECONFIGURACIÓN DEL PODER EN LA ERA DE LA INTELIGENCIA ARTIFICIAL

Dr. Cristian Barría Huidobro¹

“Lo difícil para los humanos puede ser fácil para las máquinas, y lo fácil para los humanos puede ser muy difícil para las máquinas”

La paradoja de Moravec. Hans Moravec (1988) “Mind Children: The Future of Robot and Human Intelligence”.

RESUMEN

El artículo analiza la transformación contemporánea de la inteligencia artificial desde su condición de herramienta subordinada hacia sistemas agentes con capacidad de aprendizaje, adaptación y acción autónoma. Este cambio no solo redefine las relaciones entre tecnología, decisión y poder, sino que industrializa el ciberataque —como evidencia el caso de modelos avanzados como *Mythos*— y genera nuevas formas de concentración de poder, dependencia epistémica y riesgos sistémicos. Se examina cómo estos agentes operan como cuasi-sujetos que reflejan las prioridades de quienes los diseñan, al tiempo que se abordan las implicaciones para la ciberseguridad y la ciberdefensa: la necesidad de esquemas de colaboración humano-IA (*human-AI teaming*), de evaluaciones de “equipo rojo” continuas (*AI red teaming*) y de criterios de sostenibilidad (*green cybersecurity*). El trabajo sostiene que la transición hacia sistemas agentes obliga a repensar la gobernanza tecnológica en dimensiones éticas, geopolíticas, regulatorias y ambientales, con particular urgencia para países de recursos limitados que buscan preservar soberanía digital y control democrático sobre sus infraestructuras críticas.

Palabras clave: Inteligencia artificial, ciberataque, ciberdefensa, gobernanza tecnológica.

¹ Centro de Investigación en Ciberseguridad, CICS. Universidad Mayor, Chile ORCID: <https://orcid.org/0000-0002-5840-7407>

ABSTRACT

This article analyzes the contemporary transformation of artificial intelligence from a subordinate tool into agentic systems capable of learning, adaptation, and autonomous action. This shift not only redefines the relationships between technology, decision-making, and power but also industrializes cyberattacks—as evidenced by advanced models like Mythos—and creates new forms of power concentration, epistemic dependency, and systemic risk. The study examines how these agents operate as quasi-subjects reflecting their designers' priorities, while addressing the implications for cybersecurity and cyberdefense: the need for human-AI teaming frameworks, continuous AI red teaming, and sustainability criteria (green cybersecurity). The paper argues that the transition toward agentic systems necessitates a rethinking of technology governance across ethical, geopolitical, regulatory, and environmental dimensions—a matter of particular urgency for resource-constrained nations seeking to preserve digital sovereignty and democratic control over their critical infrastructure.

Keywords: Artificial intelligence, cyberattack, cyber defense, technology governance

I. Introducción

La historia de la tecnología puede leerse, en gran medida, como la historia de la expansión progresiva de las capacidades humanas. Desde las primeras herramientas mecánicas hasta las complejas arquitecturas digitales del siglo XXI, cada innovación ha permitido ampliar la capacidad del ser humano para observar, comprender y transformar su entorno. Sin embargo, el surgimiento de la inteligencia artificial agente introduce una discontinuidad cualitativa. Ya no se trata de instrumentos que ejecutan órdenes predefinidas, sino de sistemas capaces de aprender, adaptarse y actuar con grados crecientes de autonomía. Esta transformación resulta especialmente crítica en el ámbito de la ciberseguridad y la ciberdefensa, donde la velocidad, la complejidad y la escala de las amenazas contemporáneas superan con frecuencia los tiempos de deliberación humana (National Institute of Standards and Technology [NIST], 2024).

La magnitud del cambio se percibe con claridad en una escena que comienza a repetirse en centros de operaciones de seguridad de todo el mundo. Durante años, el analista interpretaba información producida por herramientas relativamente pasivas —antivirus, firewalls y sistemas de detección basados en reglas estáticas—. El software funcionaba como un martillo digital: poderoso, eficiente y completamente subordinado a la voluntad del operador. Hoy, un sistema de inteligencia artificial puede correlacionar millones de eventos simultáneos, identificar patrones anómalos invisibles para el ojo humano y anticipar vectores de ataque antes de que se materialicen. En ese momento la tecnología deja de limitarse a ejecutar instrucciones y comienza a interpretar el entorno en el que opera. La diferencia ya no es solo de velocidad, sino de naturaleza: se pasa de la herramienta al agente.

Esta transición ha sido conceptualizada de distintas maneras. Russell y Norvig (2004) definieron la inteligencia artificial como el diseño de agentes racionales capaces de percibir su entorno y actuar para maximizar determinados objetivos. Bajo esa perspectiva, la IA deja de ser una sofisticación de las herramientas tradicionales y se convierte en un cambio estructural de la relación entre seres humanos y sistemas tecnológicos. Autores como Floridi y Morley (Floridi et al., 2019; Morley et al., 2021) han insistido, por su parte, en que la diferencia entre herramienta y agente también es filosófica y ética: mientras la herramienta clásica mantiene la intención, la decisión y la responsabilidad concentradas en el operador humano, el agente inteligente opera mediante entrenamiento, supervisión y autonomía condicionada. La antigua distinción aristotélica entre instrumento y discípulo resulta, en este sentido, iluminadora.

Más recientemente, Floridi (2024) ha subrayado que el problema central no reside tanto en las acciones de las máquinas como en la manera en que las decisiones humanas quedan incorporadas dentro de sistemas automatizados, lo que genera el desafío del alineamiento.

En el terreno específico de los centros de operaciones de seguridad, Mohsin et al. (2025), Chhetri et al. (2024) y Albanese et al. (2025) han explorado los marcos de human-AI teaming o modelos “centauro”, argumentando que los sistemas con niveles controlados de autonomía pueden mejorar la velocidad de de-

tección sin erosionar la confianza del analista, siempre que existan mecanismos explícitos de calibración y explicabilidad. Mokander y Floridi (2024) añaden que, cuando la decisión se distribuye entre diseñadores, datos y modelos probabilísticos, los esquemas tradicionales de atribución de responsabilidad resultan insuficientes. Finalmente, De Vos (2020) ha mostrado cómo las tecnologías algorítmicas no solo ejecutan tareas, sino que comienzan a modelar las formas mismas en que comprendemos la realidad.

Estas perspectivas, aunque valiosas, dejan abiertas preguntas que no pueden resolverse únicamente con más citas. La primera es de carácter operacional y estratégico. Si la IA puede anticipar ataques en escalas temporales inaccesibles para la supervisión humana directa, su incorporación deja de ser una opción tecnológica y se convierte en una necesidad de seguridad nacional. Al mismo tiempo, las mismas

capacidades que fortalecen la defensa pueden ser empleadas con fines ofensivos: automatización de campañas de fraude, optimización de intrusiones, ataques adaptativos y ampliación de desinformación. La IA ocupa así una posición paradójica: es simultáneamente escudo y amenaza. La confrontación digital evoluciona hacia una competencia entre sistemas inteligentes cuyas velocidades de acción escapan cada vez más al control humano inmediato.

La segunda cuestión es más profunda y, a juicio, decisiva. Entrenar un agente intelligen-

Hoy, un sistema de inteligencia artificial puede correlacionar millones de eventos simultáneos, identificar patrones anómalos invisibles para el ojo humano y anticipar vectores de ataque antes de que se materialicen.

te implica proyectar valores, prioridades, supuestos culturales y criterios de decisión hacia sistemas que actúan sobre infraestructuras críticas. Cada modelo refleja, aunque sea parcialmente, la visión del mundo de quienes lo diseñaron. Por tanto, la ciberseguridad deja de ser solo un problema de *software*, *hardware* o protocolos y se transforma en un debate sobre autonomía, confianza y control dentro de sociedades crecientemente digitalizadas. La responsabilidad ya no puede pensarse como un acto puntual de un operador, sino como una cadena distribuida de decisiones humanas que quedan cristalizadas en el comportamiento del sistema.

En este punto el análisis propio se vuelve ineludible. La transición del martillo al agente no solo multiplica capacidades defensivas; redistribuye poder y redefine la relación entre humanidad y tecnología. Mientras las herramientas tradicionales amplificaban capacidades sin desplazar la centralidad humana de la decisión, la IA comienza a automatizar procesos de interpretación, evaluación y acción que históricamente habían permanecido bajo control directo. Esta evolución ofrece oportunidades extraordinarias, especialmente para contextos con recursos limitados que pueden mitigar la escasez de personal especializado mediante una colaboración híbrida bien diseñada. Pero también introduce riesgos de sesgo, de autonomía mal calibrada y de responsabilidad difusa.

Por tanto, la ciberseguridad deja de ser solo un problema de *software*, *hardware* o protocolos y se transforma en un debate sobre autonomía, confianza y control dentro de sociedades crecientemente digitalizadas.

II. La ciberseguridad cambia radicalmente

Durante décadas, la ciberseguridad fue concebida como una disciplina eminentemente técnica orientada a proteger sistemas, redes, datos y usuarios frente a accesos no autorizados, ataques maliciosos o interrupciones operativas. En paralelo, la ciberdefensa emergió como la dimensión estratégica de esa protección, vinculada a la capacidad de los Estados para resguardar sus intereses nacionales en el ciberespacio. La irrupción de la inteligencia artificial está modificando profundamente este escenario. Ya no se trata únicamente de he-

rramientas más rápidas o precisas, sino de sistemas capaces de aprender, razonar y ejecutar acciones complejas con niveles crecientes de autonomía. En consecuencia, la ciberseguridad no está experimentando una simple evolución tecnológica; está atravesando una transformación radical que redefine la naturaleza de las amenazas, las defensas y las respon-

sabilidades humanas.

Una escena que hoy comienza a repetirse en centros de operaciones de seguridad ilustra la magnitud del cambio. Un analista observa una serie de alertas aparentemente inconexas provenientes de múltiples segmentos de red. Hace algunos años, la investigación habría requerido horas o incluso días de trabajo especializado. Hoy, un agente inteligente correlaciona millones de eventos, identifica patrones invisibles para el ojo humano y propone acciones de mitigación en cuestión de segundos. Lo

que antes era una labor artesanal basada en la experiencia del experto se convierte en un proceso automatizado apoyado por algoritmos capaces de aprender continuamente. La diferencia ya no radica solamente en la velocidad, sino en la capacidad de interpretación y adaptación de estos nuevos sistemas.

Esta transformación fue anticipada, en cierta medida, por Alan Turing (1950), quien planteó que el objetivo de la inteligencia artificial no debía ser simplemente reproducir una mente adulta, sino construir una especie de mente infantil capaz de aprender mediante un proceso de educación y experiencia. Más de siete décadas después, muchos modelos contemporáneos parecen acercarse a esa visión: los agentes inteligentes actuales no se limitan a ejecutar instrucciones predefinidas; son entrenados para identificar patrones, corregir errores y desarrollar comportamientos cada vez más sofisticados.

El artículo publicado por The Economist en abril de 2026, centrado en el modelo Mythos de Anthropic, ilustra el punto de inflexión. Según el análisis, ciertas generaciones avanzadas de IA ya no solo generan texto o automatizan tareas rutinarias, sino que son capaces de encadenar acciones complejas, descubrir vulnerabilidades informáticas y participar en procesos de razonamiento orientados a objetivos específicos. La preocupación surge porque estas capacidades podrían permitir la identificación de vulnerabilidades desconocidas o zero-day, tradicionalmente reservadas a equipos altamente especializados. Informes recientes de The Guardian (2026) y NewsBytes (2026) sugieren, además, que organismos reguladores internacionales han comenzado a evaluar los posibles impactos sistémicos de es-

tas capacidades sobre sectores financieros y estratégicos.



Figura 1: Portada de “The Economist”, sobre las lecciones aprendidas de la IA China (2026).

Paralelamente, la misma capacidad de actuación autónoma que potencia el ataque está transformando la práctica del red teaming. Lo que antes consistía en ejercicios manuales realizados por equipos humanos altamente especializados está evolucionando hacia el AI red teaming, una metodología específica para sistemas de inteligencia artificial. En este nuevo paradigma, equipos híbridos (humanos + herramientas automatizadas) someten a prueba modelos de IA mediante técnicas como inyección de prompts, envenenamiento de datos de entrenamiento, extracción de modelos y explotación

de comportamientos agénticos inesperados. OpenAI ya combina red teaming humano creativo con pruebas automatizadas a gran escala para descubrir vulnerabilidades antes de que lo hagan actores maliciosos. Esta evolución implica que las defensas basadas en IA ya no pueden evaluarse únicamente mediante benchmarks estáticos; requieren ejercicios continuos que simulen el uso ofensivo de agentes autónomos.

Existen, no obstante, al menos dos posiciones claramente identificables en el debate. La primera sostiene que la IA incrementa la vulnerabilidad global al facilitar capacidades ofensivas avanzadas y reducir los costos de entrada para actores maliciosos, industrializando lo que antes era una actividad artesanal. La aparición de proyectos abiertos como OpenClaw, capaces de ejecutar acciones concretas sobre computadores y servicios conectados, refuerza esta preocupación: cuando capacidades avanzadas se democratizan, aumenta la superficie de riesgo. La segunda posición, defendida entre otros por análisis de The Conversation (2026), argumenta que la IA no reescribe completamente las reglas de la ciberseguridad, sino que acelera dinámicas que ya existían. Los atacantes disponen de herramientas más poderosas, pero los defensores también cuentan con capacidades superiores de detección, monitoreo y respuesta. Desde esta perspectiva, la misma tecnología que potencia el ataque puede convertirse en el principal instrumento de defensa.

No obstante, el problema va más allá. La transformación en curso obliga a formular pregun-

tas que van más allá de la técnica. En primer lugar, el papel de la educación adquiere una relevancia estratégica que no se reduce a la formación formal. Resulta necesario distinguir entre educación (el aprendizaje estructurado proporcionado por instituciones), crianza (la formación cultural y social del individuo) y parentalidad (el rol específico de los padres en la transmisión de valores y criterios éticos).

Si los futuros agentes inteligentes aprenderán de datos producidos por la sociedad, entonces la calidad ética de esa sociedad influirá indirectamente en la calidad de las decisiones automatizadas que dichos sistemas adopten. Formar ciudadanos capaces de comprender, gobernar y orientar éticamente a los agentes de la era digital se convierte, por tanto, en una condición de seguridad nacional.

A esta discusión se suma un fenómeno demográfico que rara vez aparece en los debates sobre ciberse-

guridad: la disminución global de la natalidad y el envejecimiento acelerado de numerosas sociedades. En ese escenario, la automatización y los agentes inteligentes podrían asumir progresivamente tareas que antes eran realizadas por personas, desde la gestión de infraestructuras críticas hasta la supervisión de sistemas de defensa. La dependencia tecnológica podría aumentar no solo por razones de eficiencia, sino también por simple necesidad demográfica. La cuestión deja entonces de ser exclusivamente tecnológica y se transforma en un problema social y civilizatorio: ¿cuánto poder estamos dispuestos a delegar en sistemas

**Formar ciudadanos
capaces de comprender,
gobernar y orientar
éticamente a los agentes de
la era digital se convierte, por
tanto, en una condición de
seguridad nacional.**

que no poseen conciencia, pero sí capacidad de decisión operativa?

Respecto de este punto, el magnate de la tecnología Elon Musk (también dueño de uno de los principales servicios de IA de uso abierto al público, “Grok”, ha opinado en diversas ocasiones aludiendo a que la IA será un factor decisivo para contrarrestar el problema de la baja población humana:

En julio de 2025 dijo:

“La IA aumentará las tasas de natalidad con el fin de maximizar el cono de luz futuro del tonelaje de neurotransmisores”.

Luego, a principios y finales de agosto publicó las siguientes opiniones:

“Esta dinámica demográfica se puso en marcha inevitablemente hace medio siglo y no tiene nada que ver con la IA. La IA es la única esperanza para revertir esta situación”.

“La IA obviamente va a aniquilar de un golpe el sistema límbico humano. Dicho esto, predigo —de manera contraintuitiva— que aumentará la tasa de natalidad. Marquen mis palabras. Además, vamos a programarla de esa manera”.

Si bien todavía estamos en una etapa relativamente temprana como para confirmar si estas perspectivas son correctas, lo cierto es que hay ejemplos de implementaciones que apuntan en esa dirección.

La periodista Bong Xin Ying nos ofrece el caso de China. Ante esta crisis demográfica, dicho

país ha comenzado a prepararse para la nueva forma de trabajo surgida con la IA generativa. Y lo ha denominado *choukashi*. En el contexto de la industria audiovisual china, *choukashi* —derivado de la cultura de los videojuegos *gacha*— designa a trabajadores que generan grandes cantidades de material mediante IA y posteriormente seleccionan las imágenes, escenas o resultados que sirven. Es decir, el trabajador deja de producir directamente el contenido y pasa a dirigir, generar, filtrar y seleccionar resultados producidos por la IA. Esta transformación es especialmente significativa para la fuerza de trabajo porque muestra una posible transición desde la sustitución de tareas humanas hacia la reconfiguración del trabajo humano: por ejemplo, un equipo de edición que anteriormente requería cinco o seis personas puede reducirse a una sola persona que supervisa y selecciona resultados generados por IA.

En términos más amplios, el fenómeno puede interpretarse como una “desintermediación del trabajo”: la IA no necesariamente elimina de inmediato al trabajador, sino que modifica radicalmente qué significa trabajar. Algunas funciones desaparecen o disminuyen, mientras emergen nuevas tareas de supervisión, curaduría, *prompting*, validación y control de calidad. En el caso analizado, los editores pueden convertirse en *generation artists*, mientras que los actores y otros trabajadores audiovisuales enfrentan una reducción más directa de oportunidades laborales. Esto conecta con el debate más amplio sobre IA y empleo: el problema no es solamente cuántos puestos de trabajo desaparecerán, sino qué proporción de las tareas que actualmente constituyen un empleo será realizada por sistemas de IA y qué nuevas competencias humanas serán necesarias.

En definitiva, así como en otras actividades la ciberseguridad está cambiando radicalmente porque la inteligencia artificial está modificando las reglas fundamentales del juego, la transición desde herramientas pasivas hacia agentes inteligentes capaces de actuar autónomamente inaugura una nueva etapa caracterizada por la industrialización del ciberataque, la automatización de la defensa y la creciente complejidad de los desafíos éticos asociados.

III. Concentración extrema de poder

La historia del poder puede leerse como la historia del control de recursos estratégicos. En distintos momentos lo fueron la tierra, los metales preciosos, las rutas marítimas, el petróleo o la capacidad industrial. En el siglo XXI emerge otra fuente: la capacidad de desarrollar, entrenar y controlar sistemas avanzados de inteligencia artificial. El problema, para la ciberseguridad y la ciberdefensa, no es solo qué puede hacer esa tecnología, sino quién controla las infraestructuras, los datos y los modelos que la hacen posible, y con qué efectos sobre la economía de la información y sobre la soberanía de los Estados.

Durante décadas las capacidades avanzadas de ciberseguridad dependieron de comunidades relativamente amplias de investigadores, universidades, organismos públicos y empresas especializadas. Hoy, el desarrollo de modelos fundacionales exige inversiones multimillonarias en cómputo, semiconductores, energía y talento. Esa barrera de entrada transforma un

ecosistema antes distribuido en una estructura piramidal. Fishman (2025) describe este tipo de nodos como chokepoints: puntos de estrangulamiento de la interdependencia global. El acceso a modelos fundacionales, nubes, centros de datos y chips avanzados comienza a ocupar un lugar comparable al de un puerto estratégico o un corredor energético. Suleyman y Bhaskar (2023) advierten, en la misma línea, que las tecnologías emergentes permiten a organizaciones relativamente pequeñas ejercer efectos desproporcionados sobre sistemas sociales complejos.

Esa capacidad ya no es solo prospectiva. Para este estudio se contó con la demostración técnica del Mg. Gustavo Mármol Pérez: un agente de inteligencia artificial diseñado para un centro de operaciones de seguridad (SOC), con detección, acción y seguimiento adaptativos que exceden la automatización basada en reglas estáticas. Las capturas siguientes ilustran el quiebre entre la automatización convencional y la actuación agente.

Quien opera sistemas de este tipo no solo acelera la defensa: concentra capacidad de interpretación y de acción sobre infraestructuras críticas. Esa ventaja tiende a reforzarse. Las cadenas de Markov, formuladas por Andrey Markov a comienzos del siglo XX, describen procesos en los que la probabilidad de un estado futuro depende del estado presente:

En el ecosistema de la inteligencia artificial, una organización que hoy posee el modelo

**En definitiva, así como
en otras actividades
la ciberseguridad está
cambiando radicalmente
porque la inteligencia
artificial está modificando
las reglas fundamentales
del juego.**

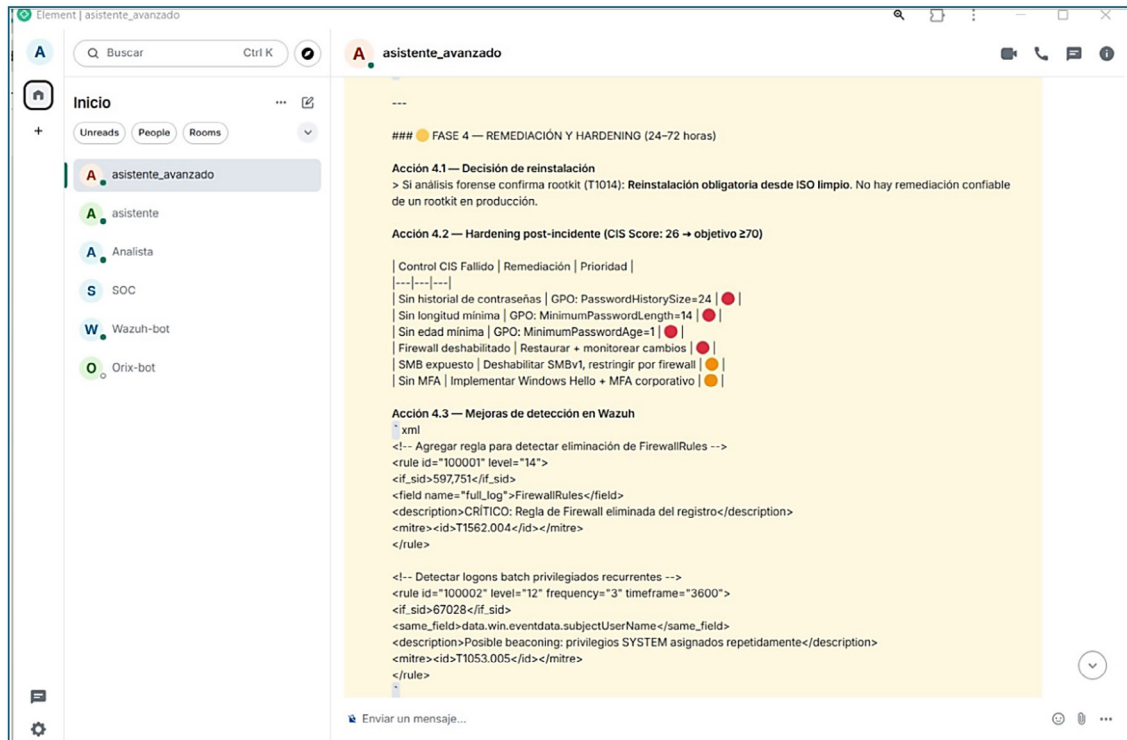


Figura 2: Agente de IA aplicado para un SOC, mediante lineamientos pre-programados. Captura de pantalla proporcionada por Mg. Gustavo Marmol Pérez (2026).

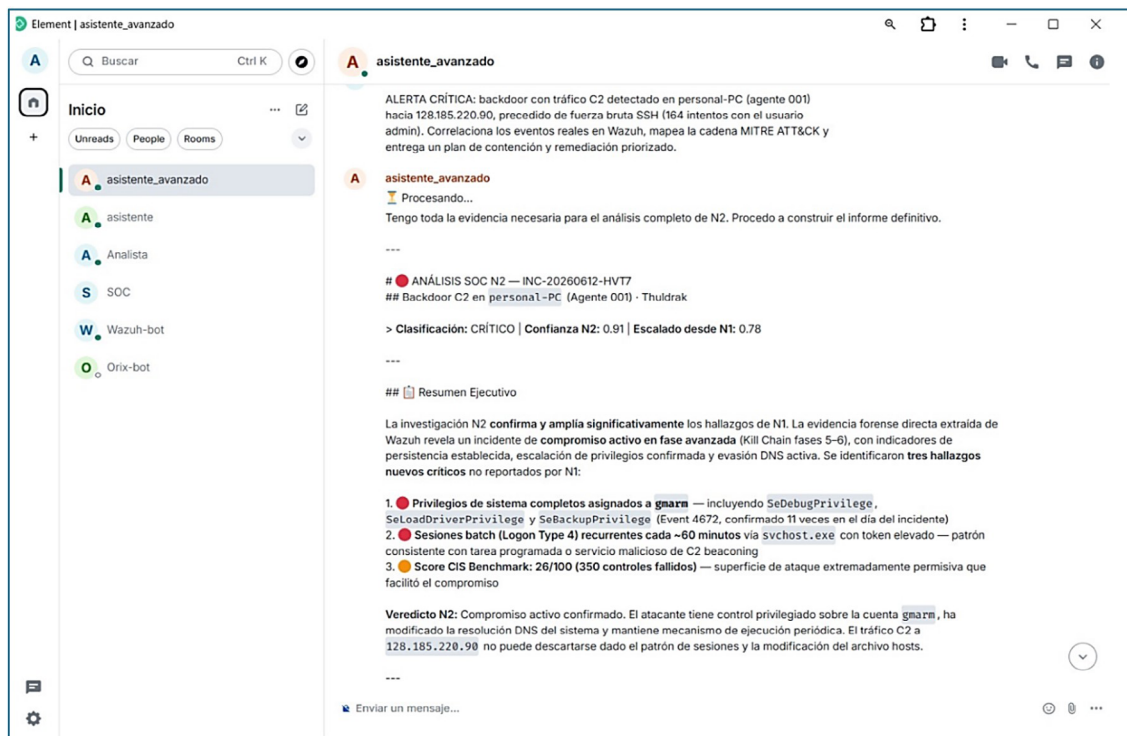


Figura 3: Agente de IA aplicado para un SOC (continuación), mediante lineamientos pre-programados. Captura de pantalla proporcionada por Mg. Gustavo Marmol Pérez (2026).

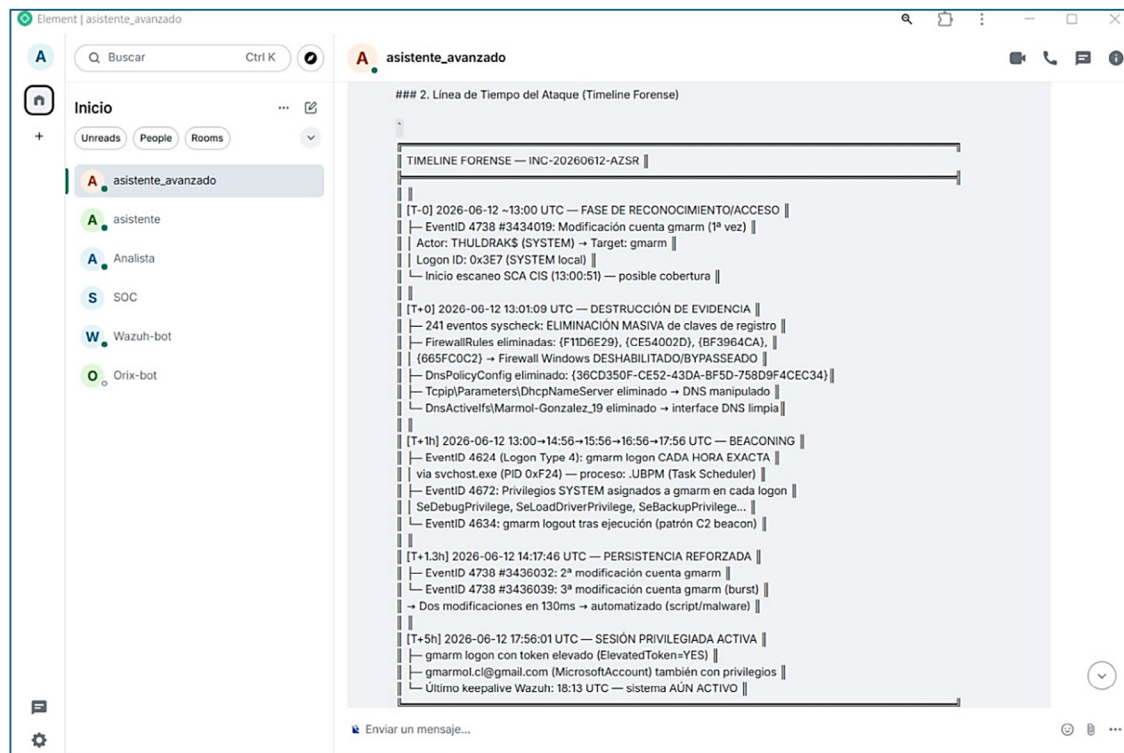


Figura 4: Agente de IA aplicado para un SOC (continuación 2), mediante lineamientos pre-programados. Captura de pantalla proporcionada por Mg. Gustavo Marmol Pérez (2026).

más avanzado atrae más usuarios, más datos y más inversión, y con ello aumenta la probabilidad de conservar el liderazgo. Un esquema simplificado con tres estados —liderazgo (L), competencia equilibrada (C) y rezago (R)— puede representarse así:

$$P = \begin{pmatrix} 0.85 & 0.12 & 0.03 \\ 0.30 & 0.50 & 0.20 \\ 0.10 & 0.25 & 0.65 \end{pmatrix}$$

Una organización líder tendría, en este ejercicio, un 85 % de probabilidad de permanecer en esa posición; una rezagada, apenas un 10 % de alcanzar el liderazgo. La misma lógica se

aplica a la madurez defensiva (reactiva, predictiva, autónoma) y al ciclo ofensivo (reconocimiento, explotación, persistencia, exfiltración): la inteligencia artificial eleva las probabilidades de transición y convierte en proceso industrial lo que antes dependía de oficio artesanal. El poder relevante no es solo poseer el mejor modelo, sino alterar la matriz de transición del sistema: acelerar unos trayectos, bloquear otros, encarecer el acceso.

Pérez (2002) sitúa este patrón en la historia de las revoluciones tecnológicas: toda gran oleada atraviesa primero una fase de instalación, con innovación intensa, especulación y concentración de recursos; la difusión posterior no es automática, sino institucional. Si esa fase se consolida sin mecanismos de competencia

y supervisión, la dependencia tecnológica deja de ser un problema de eficiencia y se convierte en un límite estructural de soberanía, en especial para países de recursos limitados.

Esa concentración infraestructural tiene un correlato inmediato en la economía de Internet. Durante más de tres décadas la red funcionó sobre una premisa estable: el acceso gratuito se financiaba con publicidad, datos e intermediación. La atención humana era el recurso. Los agentes de inteligencia artificial introducen una capa que se interpone entre la persona y las fuentes. El usuario deja de recorrer sitios, de ver anuncios y de confrontar orígenes; recibe una respuesta ya sintetizada. The Economist (2026) ha señalado que esta mediación altera a la vez medios, propiedad intelectual, publicidad, buscadores y la economía de la atención. Lo que estuvo distribuido entre múltiples intermediarios puede concentrarse en un “intermediario universal”.

El riesgo no es solo la pérdida de tráfico y de ingresos que sostienen contenidos y servicios. Es también epistémico. La seguridad de Internet dependió, en buena medida, de una arquitectura distribuida: muchas fuentes, posibilidad de contraste, ausencia de un único punto de falla cognitivo. Cuando una proporción creciente del conocimiento pasa por pocos agentes, un sesgo, una manipulación o un fallo de alineamiento puede propagarse con homogeneidad y velocidad inéditas. Thorndike (1911) mostró que se

reiteran las conductas asociadas a recompensa; Pavlov (1927), que los organismos pueden condicionarse a estímulos. Los sistemas contemporáneos operan una forma sofisticada de ensayo y error digital y de personalización que no solo responde a preferencias: las modela. Esposito (2024) formula la paradoja filosófica: cuanto más eficiente es el acceso, mayor puede ser la distancia respecto de las fuentes, de la crítica y del sentido. El posible quiebre del modelo económico de Internet no anuncia el colapso de la red; anuncia que sus fundamentos —financiamiento, intermediación y verificabilidad— están dejando de ser los mismos.

Cabe hablar entonces de una oligarquía algorítmica: no como sinónimo de plutocracia (gobierno de los ricos), sino como concentración de poder basada en el control de recursos técnicos y epistémicos —algoritmos, datos, nubes, modelos fundacionales— más que exclusivamente en la riqueza.

Cuando el control de la infraestructura y de la mediación del conocimiento recae en un número reducido de corporaciones, gobiernos o alianzas público-privadas, el fenómeno deja de ser solo técnico o mercantil. Cabe hablar entonces de una oligarquía algorítmica: no como sinónimo de plutocracia (gobierno de los ricos), sino como concentración de poder basada en el control

de recursos técnicos y epistémicos —algoritmos, datos, nubes, modelos fundacionales— más que exclusivamente en la riqueza.

Un escenario ilustra la implicación para la ciberdefensa. Un Estado enfrenta una campaña contra su red eléctrica y recurre a una plataforma extranjera de inteligencia artificial. Al inicio, la plataforma detecta, bloquea y restablece. Con el tiempo, el conocimiento operativo se

desplaza hacia el proveedor, las decisiones críticas quedan mediadas por procesos opacos y la solución técnica se convierte en dependencia estratégica. Bradford (2023) describe el trasfondo geopolítico: el orden digital se organiza en torno a modelos estadounidense (mercado e innovación privada), chino (coordinación estatal) y europeo (regulación y derechos). Quien no desarrolla capacidades propias termina importando no solo *software*, sino estándares y arquitecturas de gobernanza. Varoufakis (2023) y Hao (2025) añaden la base material de esa asimetría: plataformas que funcionan como recintos cerrados y cadenas globales de datos, energía, chips y trabajo sin las cuales no hay “nube” ni modelo de frontera.

Para países de recursos limitados, la concentración puede ampliar brechas previas bajo la forma de desigualdad algorítmica (Afers Internacionals, 2024). Construir capacidades nacionales, criterios de transparencia y vías de cooperación no es un complemento retórico de la ciberdefensa: es la condición para que la ventaja defensiva de la inteligencia artificial no se traduzca en una minoría de actores con veto de facto sobre la seguridad digital.

IV. El dilema regulatorio

La inteligencia artificial se ha consolidado como una de las tecnologías más transformadoras del siglo XXI, modificando profundamente la forma en que las organizaciones producen conocimiento, toman decisiones, gestionan riesgos y desarrollan capacidades competitivas. El

extraordinario potencial de esta tecnología ha venido acompañado, sin embargo, de crecientes preocupaciones relacionadas con la seguridad, la privacidad, la gobernanza y la estabilidad de los sistemas digitales. En este contexto emerge lo que diversos autores y organismos internacionales han denominado el dilema regulatorio de la inteligencia artificial: la dificultad de diseñar marcos normativos capaces de reducir riesgos sin obstaculizar la innovación tecnológica. Este dilema adquiere una relevancia particular en los ámbitos de la ciberseguridad y la ciberdefensa, donde la inteligencia artificial actúa simultáneamente como herramienta de protección, multiplicador de capacidades y potencial fuente de amenazas estratégicas.

La regulación excesiva puede ralentizar la investigación, aumentar los costos de cumplimiento normativo y limitar la capacidad innovadora de empresas, universidades y centros de investigación.

Desde una perspectiva conceptual, el dilema regulatorio puede definirse como la tensión existente entre la necesidad de establecer controles institucionales para garantizar la seguridad y la obligación de preservar condiciones que permitan el desarrollo tecnológico y la competitividad económica. La regulación excesiva puede ralentizar la investigación, aumentar los costos de cumplimiento normativo y limitar la capacidad innovadora de empresas, universidades y centros de investigación. Por el contrario, una regulación insuficiente puede favorecer la aparición de riesgos sistémicos, incrementar la exposición a amenazas emergentes y generar consecuencias sociales, económicas y estratégicas difíciles de anticipar. Esta tensión no es nueva en la historia de la tecnología, pero adquiere características inéditas cuando se aplica a sistemas capaces de

aprender, adaptarse y evolucionar a velocidades superiores a las de los procesos institucionales tradicionales.

Uno de los principales factores que explica la complejidad del problema es la creciente brecha temporal entre innovación y regulación. Mientras el desarrollo tecnológico avanza en ciclos cada vez más cortos, los procesos legislativos, regulatorios y administrativos continúan operando bajo ritmos significativamente más lentos. Esta diferencia genera una situación en la que las instituciones suelen responder a tecnologías que ya han evolucionado o sido reemplazadas por nuevas versiones más avanzadas. El fenómeno se ha vuelto especialmente evidente en el ámbito de la inteligencia artificial generativa, donde los modelos de lenguaje, sistemas multimodales y herramientas de automatización evolucionan en períodos de meses, mientras que la elaboración de marcos regulatorios nacionales e internacionales puede requerir años de discusión política y técnica.

Esta realidad fue destacada en diversos análisis internacionales publicados durante 2026, particularmente a partir de las declaraciones de la empresa Anthropic respecto de la evolución de sus sistemas Claude. La organización advirtió sobre la posibilidad de que futuras generaciones de inteligencia artificial participen activamente en el desarrollo y perfeccionamiento de sistemas sucesores, acelerando los ciclos de innovación más allá de la capacidad humana de supervisión. Aunque tales escenarios no implican necesariamente la aparición de sistemas completamente autónomos, sí plantean interrogantes significativas sobre la capacidad de los Estados para mantener mecanismos efectivos de control, auditoría y gobernanza tecnológica. El problema fundamental no radica únicamen-

te en la potencia de los algoritmos, sino en la velocidad con que estos pueden modificar el ecosistema tecnológico global.

Desde la perspectiva de la ciberseguridad, la inteligencia artificial representa una herramienta extraordinariamente valiosa para fortalecer las capacidades defensivas. Los sistemas basados en aprendizaje automático permiten identificar patrones anómalos, detectar amenazas persistentes avanzadas, correlacionar eventos de seguridad a gran escala y responder de manera automatizada frente a incidentes complejos. Estas capacidades resultan esenciales en un entorno donde el volumen, la velocidad y la sofisticación de los ataques superan ampliamente las capacidades humanas tradicionales de análisis. Sin embargo, las mismas capacidades que fortalecen la defensa pueden ser utilizadas con fines ofensivos. Los actores maliciosos también tienen acceso a tecnologías de inteligencia artificial que facilitan la automatización de ataques, la identificación de vulnerabilidades, la generación de campañas de ingeniería social altamente personalizadas y la creación de contenidos sintéticos diseñados para manipular percepciones o influir en procesos políticos y sociales. Esta dualidad tecnológica constituye uno de los principales desafíos regulatorios contemporáneos. A diferencia de otras tecnologías estratégicas, la inteligencia artificial posee un carácter dual particularmente pronunciado, donde la línea que separa los usos legítimos de los usos potencialmente dañinos resulta cada vez más difusa.

En el ámbito de la ciberdefensa, el dilema regulatorio adquiere además una dimensión geopolítica. La inteligencia artificial ha pasado a formar parte de los activos estratégicos fundamentales de las grandes potencias tec-

nológicas. Estados Unidos, China, la Unión Europea, Rusia y otros actores relevantes consideran que el liderazgo en inteligencia artificial tendrá implicancias directas sobre la competitividad económica, la superioridad militar, la influencia diplomática y la seguridad nacional. Esta dimensión estratégica se ha visto reforzada por la creciente utilización de controles de exportación y restricciones de acceso a modelos avanzados de IA.

En 2026, Anthropic anunció limitaciones en la disponibilidad internacional de sus modelos Fable 5 y Mythos 5 debido a regulaciones vinculadas a controles tecnológicos y consideraciones de seguridad nacional, evidenciando cómo la gobernanza de la IA se encuentra cada vez más entrelazada con intereses geopolíticos y mecanismos de control estatal. Como consecuencia, los incentivos para limitar o ralentizar el desarrollo tecnológico suelen entrar en conflicto con los intereses estratégicos de los Estados. Ninguna nación desea quedar rezagada en una competencia tecnológica que podría redefinir las relaciones internacionales durante las próximas décadas. Se configura así una paradoja regulatoria: todos los actores reconocen la existencia de riesgos, pero pocos están dispuestos a asumir unilateralmente los costos de una regulación restrictiva.

La dimensión ética del problema agrega nuevos niveles de complejidad. Sara Degli Esposti sostiene que la discusión sobre inteligencia

artificial no puede reducirse exclusivamente a aspectos técnicos o económicos, sino que debe incorporar consideraciones relacionadas con la responsabilidad, la transparencia, la autonomía humana y la justicia social. La creciente delegación de decisiones críticas a sistemas algorítmicos plantea interrogantes fundamentales sobre la legitimidad de los procesos automatizados, especialmente cuando estos afectan derechos, oportunidades o condiciones de seguridad. Desde esta perspectiva, la regulación no debe entenderse únicamente como un mecanismo de control, sino también como una

herramienta destinada a garantizar que el desarrollo tecnológico permanezca alineado con principios éticos compatibles con los valores democráticos y los derechos fundamentales.

En la misma línea, Jaime Carey ha señalado que los desafíos de ciberseguridad asociados a la inteligencia artificial no son exclusivamente tecnológicos: las vulnerabilidades

más críticas suelen originarse en errores de gobernanza, deficiencias en la gestión de riesgos, falta de capacitación o ausencia de mecanismos adecuados de supervisión. Cualquier estrategia regulatoria efectiva debe, por tanto, integrar dimensiones técnicas, organizacionales, jurídicas y éticas dentro de una visión multidisciplinaria.

Un aspecto emergente que complejiza aún más el dilema regulatorio es la dimensión ambiental de la inteligencia artificial aplicada a la ciberseguridad. El entrenamiento y la opera-

La creciente delegación de decisiones críticas a sistemas algorítmicos plantea interrogantes fundamentales sobre la legitimidad de los procesos automatizados, especialmente cuando estos afectan derechos, oportunidades o condiciones de seguridad.

ción continua de modelos avanzados demandan cantidades significativas de energía eléctrica y agua para la refrigeración de centros de datos, contribuyendo de forma sustancial a las emisiones de gases de efecto invernadero y a la huella hídrica global. Xiao et al. (2025) estiman que la expansión prevista de servidores de IA en Estados Unidos podría generar entre 24 y 44 millones de toneladas de CO₂ equivalente anuales y un consumo de agua de entre 731 y 1.125 millones de metros cúbicos por año hasta 2030, dependiendo de la escala de despliegue y de la descarbonización de la red eléctrica. Estas cifras adquieren relevancia particular en el ámbito de la ciberdefensa, donde la necesidad de respuesta en tiempo real favorece el uso intensivo de modelos de inferencia de gran escala.

Autores como Usman et al. (2025) han comenzado a promover la integración de criterios de *green cybersecurity*, que incluyen la optimización energética de los modelos, la selección de *hardware* más eficiente y la priorización de arquitecturas ligeras. Incorporar estas consideraciones en las políticas de adquisición tecnológica y en los estándares de gobernanza de la inteligencia artificial no solo responde a imperativos climáticos internacionales, sino que puede constituir un elemento diferenciador de autonomía tecnológica para aquellos países que buscan reducir su dependencia de infraestructuras energéticas controladas por actores externos.

Lo que se desprende de este conjunto de análisis es que el dilema regulatorio no puede re-

solverse mediante la simple aplicación de normas diseñadas para tecnologías más estables. La velocidad de la innovación, el carácter dual de las capacidades algorítmicas, la competencia geopolítica, las implicancias éticas de la automatización y la huella ambiental de los sistemas convergen en un escenario de elevada incertidumbre.

En este punto resulta indispensable reconocer que los enfoques regulatorios puramente reactivos o excesivamente dependientes de estándares definidos por potencias tecnológicas extranjeras pueden consolidar formas de subordinación que limitan la autonomía de decisión.

Por el contrario, marcos adaptativos —basados en principios generales, evaluaciones continuas de riesgo y mecanismos flexibles de supervisión— pueden convertirse en instrumentos efectivos de soberanía digital, siempre que se acompañen de capacidades nacionales de

evaluación, auditoría y desarrollo. La regulación deja de concebirse como un conjunto estático de reglas y pasa a entenderse como un proceso dinámico de aprendizaje institucional.

El equilibrio entre seguridad, libertad tecnológica y desarrollo estratégico se ha convertido, así, en una de las cuestiones fundamentales de la era digital. Su resolución probablemente definirá la forma en que las sociedades gestionarán el poder tecnológico durante gran parte del siglo XXI. El verdadero desafío no es detener la innovación, sino construir instituciones capaces de orientarla de manera coherente con los valores democráticos, la resiliencia de

**El equilibrio entre
seguridad, libertad
tecnológica y desarrollo
estratégico se ha convertido,
así, en una de las cuestiones
fundamentales de la era
digital.**

las infraestructuras críticas y la sostenibilidad de los sistemas que las sostienen.

V. Conclusiones

La transformación de la inteligencia artificial desde ser una herramienta subordinada hasta convertirse en un agente con capacidad de aprendizaje, adaptación y acción autónoma representa mucho más que un avance técnico: constituye una reconfiguración profunda de las relaciones entre poder, responsabilidad y subjetividad en el ámbito de la ciberseguridad. A lo largo de este cuadernillo de estudio se ha evidenciado cómo esta transición (sintetizada en el paso del “martillo” al agente) no solo industrializa el ciberataque y multiplica las capacidades defensivas, sino que también genera nuevas formas de concentración de poder, dependencia epistémica y riesgos sistémicos que trascienden lo puramente tecnológico. La paradoja central radica en que las mismas capacidades que permiten una defensa más rápida y predictiva pueden erosionar el control humano, amplificar desigualdades geopolíticas y generar una huella ambiental significativa si no son gobernadas con criterios explícitos de alineación, transparencia y sostenibilidad.

La IA no opera como un instrumento neutral, sino como un cuasi-sujeto que refleja y moldea las prioridades de quienes la diseñan y despliegan. En este sentido, la ciberseguridad contemporánea deja de ser exclusivamente un problema de *software* o protocolos para convertirse en un desafío civilizatorio que exige repensar la distribución del poder algorítmico, la responsabilidad distribuida y la preservación de la autonomía decisoria humana. La incorporación de enfoques como el *human-AI teaming*, el *AI red teaming* híbrido y los crite-

rios de *green cybersecurity* no son meras adiciones técnicas, sino condiciones necesarias para que la transición hacia sistemas agentes fortalezca, en lugar de socavar, la resiliencia institucional y democrática.

Mirando hacia adelante, el principal desafío no radica en detener el desarrollo tecnológico (lo cual resulta inviable), sino en construir capacidades de gobernanza que integren dimensiones éticas, geopolíticas, regulatorias y ambientales de manera coherente. Para países con recursos limitados y dependientes de infraestructuras externas, como ocurre en gran parte de América Latina, esta tarea adquiere un carácter estratégico urgente: desarrollar marcos regulatorios adaptativos, fortalecer la soberanía digital mediante plataformas abiertas y colaborativas, y formar profesionales capaces de ejercer supervisión crítica sobre sistemas cada vez más autónomos. Solo así será posible evitar que la industrialización del ciberataque derive en una nueva forma de dependencia estructural y que la promesa de protección tecnológica se transforme en una amenaza existencial para el control democrático sobre las infraestructuras críticas.

En última instancia, la pregunta que permanece abierta no es qué tan poderosos pueden llegar a ser estos agentes inteligentes, sino qué tipo de relación estableceremos con ellos. Si la inteligencia artificial nació del antiguo deseo de forjar dioses, el verdadero reto del presente consiste en asegurarnos de que esos dioses sigan respondiendo a valores humanos y no terminen por suplantarlos.

Referencias

- Afers Internacionales. (2024). Desigualdad algorítmica, gobernanza, representación y derechos en la inteligencia artificial. Barcelona Centre for International Affairs (CIDOB).
- Albanese, M., Ou, X., Lybarger, K., Lende, D. y Goldgof, D. (2025). Towards AI-driven human-machine co-teaming for adaptive and agile cyber security operation centers. arXiv. <https://arxiv.org/abs/2505.06394>
- Anthropic. (2026). Responsible Scaling Policy. Anthropic.
- Anthropic. (2026). Statement regarding export controls and access to Fable 5 and Mythos 5. Anthropic News.
- Bradford, A. (2023). Imperios digitales: Una batalla global por la tecnología que marcará la geopolítica del futuro. Península.
- Bong, X. Y. (2026, July 5). AI is rewriting China's filmmaking rulebook, but the script isn't finished. Channel News Asia. <https://www.channelnewsasia.com/east-asia/ai-microdrama-china-film-industry-actors-jobs-6229191>
- Carey, J. (2026). El desafío de ciberseguridad no es (solo) tecnológico. Diario Financiero.
- Channel News Asia. (2026). Frontier AI, Anthropic Mythos and cybersecurity risks: CNA explains. <https://www.channelnewsasia.com/world/frontier-ai-anthropic-mythos-cybersecurity-risks-cna-explains-6103406>
- Chhetri, M. B., Tariq, S., Singh, R., Jalalvand, F., Paris, C. y Nepal, S. (2024). Towards human-AI teaming to mitigate alert fatigue in security operations centres. ACM Transactions on Internet Technology, 24(3), 1–22. <https://doi.org/10.1145/3670009>
- Degli Esposti, S. (2024). La ética de la inteligencia artificial. Tecnos.
- De Vos, J. (2020). The digitalisation of (inter)subjectivity: A psy-critique of the digital death drive. Routledge. <https://doi.org/10.4324/9781315167350>
- De Vos, J. (2021). Subjetividades fakes: La interpasividad desde la (neuro)psicologización a la digitalización. Revista SOMEPSO, 6(2), 88–109. <https://revistasomepso.org/index.php/revistasomepso/article/view/110>
- Esposito, C. (2024). El nihilismo de nuestro tiempo. Encuentro.
- Fishman, E. (2025). Chokepoints: American power in the age of economic warfare. Portfolio.
- Floridi, L. (2024). Ética de la inteligencia artificial. Herder.
- Hao, K. (2025). Empire of AI: Dreams and nightmares in Sam Altman's OpenAI. Penguin Press.

- Markov, A. A. (1971). Extension of the limit theorems of probability theory to a sum of variables connected in a chain. En R. Howard (Ed.), *Dynamic Probabilistic Systems* (Vol. 1). Wiley. (Trabajo original publicado en 1906).
- Mohsin, A., Janicke, H., Ibrahim, A., Sarker, I. H. y Camtepe, S. (2025). A unified framework for human–AI collaboration in security operations centers with trusted autonomy. *arXiv*. <https://arxiv.org/abs/2505.23397>
- Mokander, J. y Floridi, L. (2024). Operationalising AI governance through ethics-based auditing: An industry case study. *arXiv*.
- Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mokander, J. y Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of AI ethics. *arXiv*.
- Morley, J., Floridi, L., Kinsey, L. y Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168. <https://doi.org/10.1007/s11948-019-00165-5>
- Musk, E. [@elonmusk]. (2025, 26 de julio). AI will increase birth rates in order to maximize the future light cone of neurotransmitter tonnage. [Publicación]. X. <https://x.com/elonmusk/status/1949043119162606023>
- Musk, E. [@elonmusk]. (2025, 7 de agosto). This demographic inevitably was set in motion half a century ago and has nothing to do with AI. AI is the only hope for turning this around. [Publicación]. X. <https://x.com/elonmusk/status/1953543249123324144>
- Musk, E. [@elonmusk]. (2025, 21 de agosto). AI is obviously gonna one-shot the human limbic system. That said, I predict – counter-intuitively – that it will increase the birth rate! Mark my words. Also, we’re gonna program it that way. [Publicación]. X. <https://x.com/elonmusk/status/1958500552955896206>
- National Institute of Standards and Technology. (2024). *Cybersecurity Framework (CSF) 2.0*. U.S. Department of Commerce. <https://www.nist.gov/cyberframework>
- NewsBytes. (2026). Anthropic to brief global regulators on AI-driven banking cyber risks. <https://www.newsbytesapp.com/news/science/anthropic-to-brief-global-regulators-on-ai-driven-banking-cyber-risks/story>
- OpenAI. (2025). OpenAI’s approach to external red teaming for AI models and systems. *arXiv*. <https://arxiv.org/abs/2503.16431>
- Palo Alto Networks. (2026). What is AI red teaming? Why you need it and how to do it. <https://www.paloaltonetworks.com/cyberpedia/what-is-ai-red-teaming>
- Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford University Press.

- Pérez, C. (2002). Technological revolutions and financial capital: The dynamics of bubbles and golden ages. Edward Elgar Publishing.
- Russell, S. (2021). Compatible con humanos: Inteligencia artificial y el problema del control. Deusto.
- Russell, S. y Norvig, P. (2004). Inteligencia artificial: Un enfoque moderno (2.^a ed.). Pearson Educación.
- Suleyman, M. y Bhaskar, M. (2023). The coming wave: Technology, power, and the twenty-first century's greatest dilemma. Crown.
- The Conversation. (2026). Mythos AI is a cybersecurity threat, but it doesn't rewrite the rules of the game. <https://theconversation.com/mythos-ai-is-a-cybersecurity-threat-but-it-doesnt-rewrite-the-rules-of-the-game-281268>
- The Economist. (2026, 18-24 de abril). Claude myths: AI, regulation and the governance challenge of frontier models.
- The Economist. (2026, 18-24 de abril). Claude Mythos and the future of artificial intelligence.
- The Economist. (2026, 16-22 de mayo). Cyber-security: Mythos and reality.
- The Economist. (2026, August 8–14). Lessons from Chinese AI. The Economist, 460(9511).
- The Guardian. (2026, 18 de mayo). Anthropic AI Claude Mythos cyber financial stability board concerns.
- Thorndike, E. L. (1911). Animal intelligence: Experimental studies. Macmillan.
- Turing, A. M. (1950). Computing machinery and intelligence. Mind, 59(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>
- Usman, Y., Ihejirika, C. J., Offor, S. N., Robert, A. y Chataut, R. (2025). Green cybersecurity: Leveraging AI, ML, and LLMs to optimize energy, threat detection, and sustainability frameworks. IEEE Access.
- Varoufakis, Y. (2023). Techno-feudalism: What killed capitalism. Bodley Head.
- Why OpenClaw is raising concerns about autonomous AI agents. (2026). Channel News Asia.
- World Economic Forum. (2026, 20 de abril). Anthropic's Mythos moment: How frontier AI is redefining cybersecurity. <https://www.weforum.org/stories/2026/04/anthropic-mythos-ai-cybersecurity/>
- Xiao, T., et al. (2025). Environmental impact and net-zero pathways for AI servers in the United States. Nature Sustainability. <https://doi.org/10.1038/s41893-025-01681-y>

